PB-237 807-T

CONTRIBUTION TO THE INTERPRETATION OF MASS SPECTRA WITH THE AID OF ADAPTIVE LINEAR CLASSIFIERS

K. Varmuza, et al

Environmental Protection Agency
Research Triangle Park, North Carolina

1974

DISTRIBUTED BY:

NTIS

National Technical Information Service
U. S. DEPARTMENT OF COMMERCE

212036

NTIS

PB 237 807-T

EPA-TR 492-74

K. Varmuza, P. Krenmayr

CONTRIBUTION TO THE INTERPRETATION OF MASS SPECTRA WITH THE AID OF ADAPTIVE LINEAR CLASSIFIERS

[Beitrag zur Interpretation von Massenspektren mit Hilfe adaptiver, linearer Klassifikatoren]

Z. Anal. Chem., Vol. 266, No. 4, 1973, pp. 274-281.

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. Department of Commerce
Springfield, VA 22151

Translated for the
ENVIRONMENTAL PROTECTION AGENCY
by
Translation Consultants, Ltd., Arlington, Virginia

G

24

K. Varmuza and P. Krenmayr

CONTRIBUTION TO THE INTERPRETATION OF MASS SPECTRA WITH THE AID OF ADAPTIVE LINEAR CLASSIFIERS

Summary

Molecular structures of organic substances with low molecular weight can be determined with some probability from low resolution mass spectra by linear, binary, adaptive classifiers. 250 mass spectra were used as a training set to compute such classifiers for 31 different molecular structures; other 250 mass spectra were used to test the predictive ability P. The best preprocessing of the spectral data was the logarithmic transformation of the intensities and normalization to the base peak. If the training set contains only some members of a certain class, P is very low for spectra of this class and the trained classifier is not useful. After critical examination only 10 molecular structures remained, which are suitable for this method of automatic mass spectra interpretation.

Introduction

The installation of a mass spectrometer, which is coupled to a gas chromatograph and to a data system, makes it possible to obtain within a short period of time a great deal of information about the sample under study. The greatest work effort is frequently required for the interpretation of the large number of mass spectra which ensue from such an analysis. It is thus desirable to use the computer, apart from data identification and spectra preparation, also as an aid in the process of spectra interpretation. The simplest and, at the present time, most frequently utilized method consists of a procedure

in which the individual spectra are compared with a data bank, which is stored on a magentic tape and in the course of which, based on a specified similarity criterion, the substance definitions of those spectra that show the greatest degree of conformity are listed in a print-out. The disadvantages of this method consist of the necessity to work with large spectra data banks, as well as of the fact that one is limited to those substances that are part of the data bank. More pretentious computer programs with "artificial intelligence" attempt by way of heuristic methods to approximate, for purposes of spectra interpretation, the operating process of mass spectroscopic devices [2,4]. In spite of the difficulty of combining in one pattern the frequently not sufficiently known disintegration rules, which contain many exceptions, these programs have an astounding degree of efficiency; they do, however, require large computer installations which are not always available.

Another method for automatic spectra interpretation consists of identifying a mass spectrum as "pattern" which contains in coded form information concerning the chemical structure of the compound under examination. Without insight into the chemical processes in the mass spectrometer, but also free of speculative disintegration processes on the part of the ions, it is attempted to allocate

the mass spectrum to a substance category or a structure element by using the mathematical methods of automatic pattern recognition. Jurs, Kowalski, Isenhour and Reilley [7,8,10-13,16] were the first to test the application of this method for mass spectra in the form of adaptive linear classifiers (also referred to as "learning machine" or as decision vectors).

It is true that the computation of the decision vectors requires considerable computing work; the classification itself, however, can be easily undertaken with either a small-size computer or table calculating machine.

This report relates individual experiences in connection with the determination of substance categories and chemical structures of low-molecular compounds with the aid of adaptive, linear classifiers emanating from low-resolution mass spectra.

After a short introduction dealing with the method, one takes up the effect of the intensity transformation and intensity standardization upon the quality of the decision vectors. The classification characteristics, which so far have been dealt with only summarily, are critically examined, so as to delineate the boundaries within which the method can be effectively utilized.

Classification of mass spectra by means of decision vectors

The theory of adaptive, linear classifiers [18,19] and its application to the interpretation of low-resolution mass spectra [7,8,10,11,17] has been described in detail; thus a short introduction suffices at this point.

The automatic pattern recognition has the goal of classifying, on the basis of certain measured characteristics, those objects or properties which are not directly ascertainable through mechanical means (such as the classification of letters on the basis of the measured distribution of black spots in the text). Frequently, one can observe a complicated connection, not based on any theory, between the characteristics and the category association. One attempts through statistical methods to determine from a random sample of objects with specified category classification, a classifier (i.e., a calculation specification) which classifies the random sample with as low an error ratio as possible. The total of all characteristics of a random sample element is usually referred to as "pattern" and summarized in a characteristics vector. In our case, a characteristics vector corresponds to the mass spectrum of a certain substance; the classifiers are expected to use this as a basis for determining the substance category and the chemical structure.

A low-resolution mass spectrum of a substance i can be represented as spectra vector $\vec{Y}_i$ in a polydimensional space. Each dimension of this spectra space corresponds to a certain mass number; the coordinates of the vector apex, for instance, are indicated by the peak heights in the spectrum, whereas the vector proceeds from the source. If the spectra vectors of substances of a specific category are spatially separated from the other spectra vectors, it is frequently possible to find a hyperlevel which corresponds to this separation. On one side of this decision level lie the spectra vectors of substances that belong to the selected structure category (Category "YES"); on the other side lie those which do not belong to said structure category (Category "NO"). By expanding all spectra vectors through addition of another component, which may, for instance, have a value of 1, it is always possible to have the decision level cross the point of origin.

As classifier, one uses a decision vector $\vec{W}$ which emanates from the point of origin and whose course is perpendicular to the decision level. The symbol of the scalar product $S_i = \vec{W} \cdot \vec{Y}_i$ indicates to which category a substance i belongs (for instance, $S > 0$ for Category "YES," $S < 0$ for Category "NO").

For the computation of a decision vector, one needs a random sample of spectra vectors with known category association ("training set"). These spectra vectors were repeatedly introduced to an arbitrarily chosen starting vector $\vec{W}_0$. If the classification is correct, the decision vector remains unchanged; in the case of erroneous classification, the vector is corrected (adapted), for instance through reflection of the decision level at the falsely classified spectra vector, in accordance with the following formula:

$$\vec{W}' = \vec{W} - \frac{2(\vec{W} \cdot \vec{Y})}{(\vec{Y} \cdot \vec{Y})} \cdot \vec{Y} = \vec{W} - 2\frac{(\vec{W} \cdot \vec{Y})}{|\vec{Y}|^2} \cdot \vec{Y}. \qquad (1)$$

This "training process with negative feedback" continues until all spectra vectors of the training set can, without any additional vector corrections, be accurately classified ("convergence"); otherwise, it is discontinued after a certain period of time. The training process is evaluated by means of the convergence velocity. This is assessed on the basis of the number of completed classifications (N_{class}) or the number of vector corrections (N_{corr}), until the convergence is achieved. The classification capability P (predictive ability) of the obtained decision vector is measured against spectra vectors which had not been used for the training ("prediction set"). In addition, the classification capability with respect to disturbed spectra

of the training or prediction set is also of significance. As will be shown later on, P ought not to be indicated only for the total prediction set (P_{tot}), but must be listed for each category individually (P_{YES}, P_{NO}).

Data and programs

For 500 different substances with molecular weights ranging between 16 and 150, and with the chemical formula $C_{1-11}H_{1-22}O_{0-2}N_{0-2}$, mass spectra from spectra compilations [1,20] were punched. Peaks which amounted to less than 1% of the base peak were not taken into consideration. Among these 500 mass spectra, 138 different mass numbers are taken up by a total of 13,896 peaks. After mixing the 500 spectra with a disintegration generator program, 250 spectra were used as a training set and the other 250 as a prediction set. Through effective storage utilization, it was possible for the entire training set to be allocated space next to the programs in the 32K/36 bit-core memory device of the employed IBM 7040 Computer[1]. All programs were recorded in FORTRAN IV.

The training of the decision vectors was undertaken for 31 different classifications of substance categories, chemical structure elements, as well as for oxygen and nitrogen presence (Table 2).

---

[1] Computer center of the Technical University at Vienna.

The gathering of 250 spectra from the tape and the preparation of the spectra required approximately 60 seconds. The calculation period of the training program in seconds is expressed approximately by the formula

$$0.033 \cdot N_{class} + 0.075 \cdot N_{corr}$$

for instance, 140 seconds for 3,600 classifications with 300 vector corrections. The training process was carried out according to the "sub-set method" with vector correction, on the basis of formula (1) as described by [illegible]rs [8,21], after a maximum of 10,000 classifications the process was discontinued.

Results

Intensity transformation and standardization

The manner in which the characteristics (in our case, the peaks of the mass spectrum) are summarized in the characteristics vector (spectra vector), affects the cluster formation and the separability of the categories. Since the theory of automatic pattern recognition does not offer an ultimate solution for the preliminary data treatment, one is forced to rely on experimental results [18,19]. In the simplest version, each dimension of the spectra space corresponds to a mass number, and the coordinates of the vector apexes are indicated by the intensity data, which are usually linear in mass spectra,

in relation to the base peak. There are, however, also other possibilities of describing spectra vectors.

The non-linear change of the peak intensities within a spectrum is referred to as transformation. The uniform presentation of all (transformed) spectra is called standardization; it does not change the intensity conditions within the spectra, but affects the intensity parameters between the spectra. Apart from the standardization of the base peak, it is customary in mass spectrometry to standardize the total of all peaks within the spectrum. Finally, the spectra vectors may also be standardized, as far as equal length is concerned, so that the vector apexes end up by being situated on a hyper-sphere [3]. Little is known about the effect of the standardization upon the quality of the decision vectors [11].

Table 1 contains several results obtained in the course of our studying the effect of the intensity transformation and standardization, based on the classification example, upon the presence of nitrogen atoms, of a benzene ring, and of aliphatic alcohol (other substance categories showed similar results). The original intensities I (1 to 100%) were transformed in accordance with the formulas

$$I' = 10 \cdot \log (2 \cdot I) \text{ and } \quad I' = \sqrt{I} \qquad (2)$$

Table 1. Effect of transformation and standardization of the peak intensities upon the training process and upon the classification capability P_{tot} of the total prediction set; N_{class} and N_{corr}, respectively, represent the number of required classifications and vector corrections, respectively, up to the achievement of the convergence

Standardiz.	Transform.	Nitrogen			Aliphatic alcohol			Benzene ring		
		N_{class}	N_{corr}	P_{tot}	N_{class}	N_{corr}	P_{tot}	N_{class}	N_{corr}	P_{tot}
Base peak	without	1338	88	82,0	>10000	>1333	84,0	1036	36	94,0
	log	1620	120	90,4	3563	313	90,0	1310	60	95,6
	√	1356	106	85,6	3024	274	90,8	1029	29	96,0
Cone	without	2092	342	82,4	6968	968	89,6	1847	97	96,4
	log	2020	270	89,6	2925	425	91,2	2115	115	96,0
	√	2500	250	86,4	3743	493	93,2	1837	87	95,6
Total	without	3622	622	83,2	>10000	>3959	86,8	1979	229	96,0
	log	7853	2603	92,8	>10000	>5512	91,6	4954	1704	95,2
	√	4717	1717	92,4	>10000	>4783	91,2	4428	678	96,8
–	binary	2293	293	86,8	2290	290	90,0	1818	68	96,8

The standardizations were based on the vector length 1 (radius of the hyper-sphere) and on the intensity sum 1, respectively. The training was in each case begun with two different starting vectors $\vec{W}_0$ (all components +1 or -1); the classifier contained in the table is the one that possesses the greater classification capability P_{tot} for the entire prediction set. The logarithms or square roots of the peak intensities resulted, independent of the standardization, in almost all cases in a decision vector with greater P_{tot} than in the training with linear intensities. This result confirms other results for the classification of oxygen presence in the molecule [9, 11].

The standardization of the vector apexes upon a hyper-sphere did not bring about any improvements in comparison with the base peak standardization. The utilization of mass spectra, standardized with respect to the total intensity, resulted in decision vectors with a slightly improved classification capability, although at the expense of a considerably longer training duration. Further research was based on the logarithmical mass spectra with standardized base peak. In summary, the following may be said: The non-linear intensity transformation (change of vector direction) has a greater effect upon the classification capability than the change of the vector length in the course of the standardization process. The latter affects primarily the convergence velocity.

If one assumes the same value for the peaks of all mass spectra (for instance, the value 1), one obtains binary-coded spectra. In spite of the tremendous data reduction, such mass spectra still have a high information content [5,6]. As the last line in Table 1 shows, the decision vectors, determined with the aid of such spectra, possess an astonishingly high classification capability. As far as several substance categories are concerned, however, it occurred that two binary-coded spectra of different categories were identical, and the training process, of necessity, failed to function.

Table 2. Classification according to substance categories and chemical structures with binary decision vectors. (Training set and prediction set: in each case, 250 mass spectra with logarithmical intensities with standardized base peak)

No.	Category	Percentage of category in		Training-process		Class.capability(P) f.unknown spectra			Reliability(V) and no. of yes/no answers			
		Train.-set	Pred.-set	N_{class}	N_{corr}	P_{tot}	P_{YES}	P_{NO}	V_{yes}		V_{no}	
1	2	3	4	5	6	7	8	9	10	11	12	13
1	*Alkan	9,2	6,8	1797	47	96,0	88,2	96,6	65,2	(23)	99,1	(227)
2	*Cycl. KW	3,6	3,2	2100	100	96,4	75,0	97,1	46,2	(13)	99,2	(237)
3	Alken	5,6	5,2	2423	173	94,0	46,2	96,6	42,9	(14)	97,0	(236)
4	Dien	1,2	2,0	708	18	96,4	40,0	97,6	25,0	(8)	98,8	(242)
5	*Alkylbenzol	6,0	6,0	1294	44	97,2	80,0	98,3	75,0	(16)	98,7	(234)
6	*Aliph. Alkohol	12,4	11,2	3563	313	90,0	78,6	91,4	53,7	(41)	97,1	(209)
7	Äther	11,6	9,2	2528	278	91,2	47,8	95,6	52,4	(21)	94,8	(229)
8	Keton	7,6	8,8	1405	155	90,4	36,4	95,6	44,4	(18)	94,0	(232)
9	Aliph. Keton	5,6	7,2	2192	192	91,6	44,4	95,3	42,1	(19)	95,7	(231)
10	Aromat. Keton	1,2	0,8	758	8	99,2	50,0	99,6	50,0	(2)	99,6	(248)
11	Aldehyd	4,4	6,8	2172	172	89,6	23,5	94,4	23,5	(17)	94,4	(233)
12	Carbonsäure	3,2	2,8	1063	63	96,4	42,9	97,9	37,5	(8)	98,4	(242)
13	Ester	8,0	8,8	3364	364	86,4	31,8	91,7	26,9	(26)	93,3	(224)
14	Carbonylgruppe	25,6	30,4	6751	1001	78,4	64,5	84,5	64,5	(76)	84,5	(174)
15	Aliph. CO	19,6	24,8	7081	1081	78,4	54,8	86,2	56,7	(60)	85,3	(190)
16	*Aromat. CO	4,4	4,8	1047	47	97,6	83,3	98,3	71,4	(14)	99,2	(236)
17	*NH_2-Gruppe	13,6	14,8	1337	87	90,4	70,3	93,9	66,7	(39)	94,8	(211)
18	Amid	2,4	3,2	1284	34	95,2	0,0	98,4	0,0	(4)	96,8	(246)
19	Aliph. NH_2	10,8	13,2	1318	68	92,0	69,7	95,4	69,7	(33)	95,4	(217)
20	Aromat. NH_2	2,4	1,6	782	32	97,6	25,0	98,8	25,0	(4)	98,8	(246)
21	*Benzolring	23,6	21,6	1310	60	95,6	88,9	97,5	90,6	(53)	97,0	(197)
22	*Phenol	4,4	4,0	1564	64	98,4	80,0	99,2	80,0	(10)	99,2	(240)
23	Pyridinring	8,0	6,0	1323	73	94,8	66,7	96,6	55,6	(18)	97,8	(232)
24	Methylgruppe	77,2	78,0	6958	1458	80,4	85,1	63,6	89,3	(186)	54,7	(64)
25	Äthylgruppe	37,2	43,2	> 10000	> 2769	68,0	58,3	75,4	64,3	(98)	70,4	(152)
26	n-Propylgruppe	20,8	23,6	7286	1536	76,0	37,3	88,0	48,9	(45)	82,0	(205)
27	Vinylgruppe	10,0	9,6	3707	457	86,4	25,0	92,9	27,3	(22)	92,1	(228)
28	C=C-Doppel-bindung	16,8	19,6	6016	766	82,4	42,9	92,0	56,8	(37)	86,9	(213)
29	Verzweig. am C	23,6	25,2	4618	618	81,2	57,1	89,3	64,3	(56)	86,1	(194)
30	*Sauerstoff	52,0	54,8	1446	196	83,6	86,1	80,5	84,3	(140)	82,7	(110)
31	*Stickstoff	25,2	24,0	1620	120	90,4	83,3	92,6	78,1	(64)	94,6	(176)

Key to column 2:

1 *Alkane
2 *Cyclic KW
3 Alkenes
4 Dienes
5 *Alkylbenzol
6 *Aliphatic alcohol
7 Ether
8 Ketone
9 Aliphatic ketone
10 Aromatic ketone
11 Aldehyde
12 Carbonic acid
13 Ester
14 Carbonyl groups
15 Aliphatic CO
16 *Aromatic CO
17 *NH_2-group
18 Amide

19 Aliphatic NH_2
20 Aromatic NH_2
21 *Benzene ring
22 *Phenol
23 Pyridine ring
24 Methyl group
25 Ethyl group
26 n-propyl group
27 Vinyl group
28 C=C-double bond
29 Branching at C
30 *Oxygen
31 *Nitrogen

Classification according to substance categories and chemical structure elements

The classification of mass spectra may be achieved through several independent decision vectors, each of which gives either a "yes" or "no" answer to a question concerning the presence of a substance category or a chemical structure [13]. For 31 such questions, two decision vectors each were trained with the logarithmical, base-peak standardized spectra of the training set (the components of the starting vector $\vec{W}_0$ were fixed at +1 or -1). The values in Table 2 refer in each instance to that vector which possesses the greatest classification capability for the entire prediction set. Only in classifying the presence of (at least) one ethyl group, did it prove impossible prior to the time the training was discontinued to achieve convergence in the case of 10,000 completed classifications.

The category designations in column 2 have the usual chemical meaning. Categories 1 to 5 do not contain any heteroatoms. Category 7 contains also cyclic ether, but no epoxides. Category 9 contains only pure aliphatic ketones; Category 10 also contains mixed aliphatic-aromatic ketones.

Category 14 contains all substances with a CO-grouping; it is, moreover, subdivided into aliphatic (Category 15) and aromatic (Category 16) compounds. Amines, amides, and the corresponding substituted compounds are contained in Category 17; for the distribution in Category 19 and in Category 20, the same applies to the CO-group. In Category 28, all compounds with C=C double-bond, except for the aromatics, are combined. Category 29 contains substances that possess a C-atom which is linked to, at least, three other C-atoms. The percentages of each category in the training set and in the prediction set are listed in columns 3 and 4 of Table 2.

The classification capability P_{tot} for the entire prediction set is, in almost all decision vectors, very good and generally supplies an accurate classification for more than 90% of the spectra. Similarly good results were obtained by Jurs and collaborators [7,13]. In actual practice, however, these percentages are of only minor significance, since the distribution of the spectra into the separate "YES" and "NO" categories is frequently very irregular. One ought to bear in mind the following: If, for instance, only 10% of the spectra of the prediction set belong to the "YES" category, a primitive classifier,

which constantly answers "no", would accurately evaluate 90% of the unknown spectra. A good classifier, however, ought to accurately classify a high percentage of both categories. Among the 31 question formulations, which were being studied, there were only ten cases for which the decision vector accurately classified the "YES" category (P_{YES}), as well as the "NO" category (P_{NO}), at a rate of at least 70%. The user of a classifier views the reliability problem also from another point of view. He would like to know with what degree of probability the answer provided by the classifier may be considered to be correct. Columns 11 and 13 of Table 2 contain the number of the "yes" and "no" answers obtained during the classification process of the prediction set. The percentage of correct answers is listed in Columns 10 and 12. In most instances the reliability rate of the no-answer (V_{no}) is very high, whereas for the yes-answer (V_{yes}) it amounts to only 50% or less. Such decision vectors are worthless for practical purposes. This inadequacy of the linear, adaptive classifiers for the interpretation of spectra has so far not been clearly established. As is easily conceivable, the problem in question is of a fundamental nature: Since classifiers are needed for the greatest possible number of questions concerning substance categories and chemical

structures, each spectra collection, and thus the training set, contains in most cases only few spectra out of each category. A low number of spectra in a category (most frequently in the category "YES") does permit a rapid convergence during the training process; however, the decision vector thus obtained has frequently little relation to the classification problem. If one needs a reliability rate of at least 70% for the yes/no-answers, the number of substance categories to which the method can be effectively applied is still further reduced. The present report shows good results only for the following classifications: Presence of oxygen; benzene ring in the molecule; presence of alkylbenzol, of phenol, or of an aromatic compound with CO-grouping.

In this context it was analyzed how the training progresses if the spectra are arbitrarily divided into two categories. On the basis of theoretical deliberations [18,19], up to 2 n vectors, with n components each, are linearly separable in any desired distribution. The partition of the 250 spectra of the training set was undertaken by means of a disintegration generator program; the share of the spectra in the "YES" category fluctuates between 4% and 96%. The training experiments were carried out with the same training set as before and were discontinued upon completion of a maximum of 10,000 classifications.

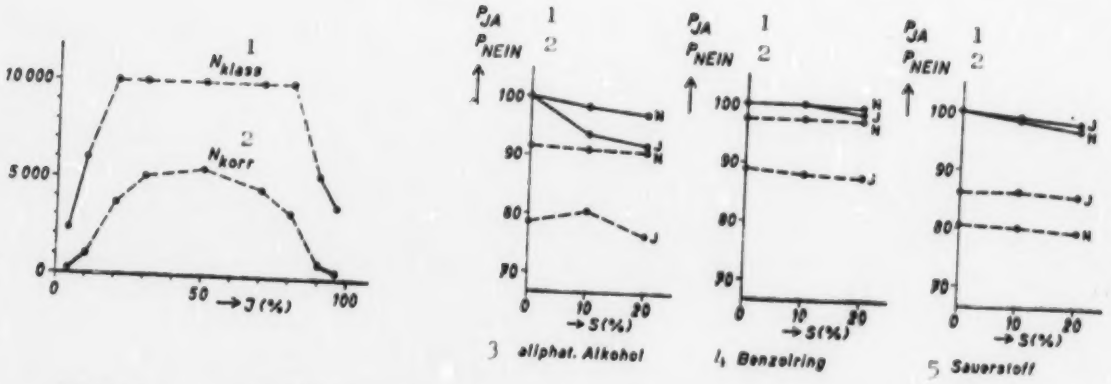

Figure 1. Training experiments with 250 mass spectra at arbitrary category classification (o-o convergence, 0--0 no convergence up to 10,000 classifications; N_{class} and N_{corr} are mean values emanating from two different training experiments; J represents the share of the spectra in the "YES" category)

Key: 1-N_{class}; 2-N_{corr}

Figure 2. Classification capability of the faded training set (——) and prediction set (---). P_{YES} and P_{NO} are the percentages of the accurately classified spectra from the "YES" and "NO" categories. S represents the standard deviation.

Key: 1-P_{YES}; 2-P_{NO}; 3-Aliphatic alcohol; 4-Benzene ring; 5-Oxygen.

As Figure 1 shows, no convergence occurred in those cases in which more than 10% of the spectra belongs to one category. In the case of less than 10%, the convergence velocity at arbitrary category distribution was noticeably lower than in the case of spectra distribution according to chemical aspects. This indicates that, nevertheless, a certain cluster formation exists. The indistinctly indicated clusters are, however, so poor in spectra that by employing the heretofore described method it is not possible to calculate an effective classifier.

Table 3. Average values of the classification capability of 10 vectors in relation to faded spectra

Spectra set	Category	Percentage of accurately classified spectra in the course of intensity fading with the standard deviation		
		0%	10%	20%
Training set	"YES"	100.0	96.9	95.5
	"NO"	100.0	99.4	98.8
Prediction set	"YES"	81.4	81.8	80.1
	"NO"	94.5	94.4	94.2

It is a known fact that the mass spectra of many substances are not easily reproducible. For this reason, the classification capability of the decision vectors in relation to disturbed mass spectra is also of significance. The disturbance of the spectra occurred due to the fading of the peak intensities. They were statistically changed for each program in such a manner that the percentage deviations from the original intensities correspond to a Gauss distribution with a standard deviation of 10% or 20%, respectively. The maximum deviation from the original intensity was limited to thrice the standard deviation. Subsequently, the spectra were again standardized to agree with the base peak. Each spectrum was faded four times; this resulted in 1,000 different spectra each for the faded training and prediction sets. These spectra were classified with the ten best decision vectors, which are marked with an asterisk in Table 2.

Several characteristic results are depicted in Figure 2. The decline in classification capability with respect to the original spectra is greater in the training set than in the prediction set, since in the latter the fading is frequently not accompanied by any deterioration. In Table 3 the average values for all 10 classifications are summarized. They show the remarkably high classification capability of the decision vectors for faded spectra.

Discussion

Several advantages and disadvantages of the adaptive, linear classifiers for the interpretation of mass spectra or other chemical data have already been thoroughly discussed [7,14,15]. The results of the present research have shown that the method in the described simple form is suitable only for the classification of few substance categories or chemical structures. The spectra collections used up to now contain the greatest possible number of substance categories. The numerical division of the spectra into "YES" and "NO" categories is thus, of necessity, frequently very irregular, and the resulting decision vectors have little relation to the chemical classification problem. In these cases the question, for the time being, remains unanswered as to whether the relatively small number of spectra of a certain substance category in a large spectra quantity actually contain enough information for classifying

purposes. If the spectra contain such information -- training experiments with spectra, which were divided into arbitrary categories, indicated that such might be the case -- it may be possible for other methods of automatic pattern recognition to be successful under such conditions as well.

As far as the small-size computers are concerned, which have already found widespread acceptance, the interpretation of mass spectra or other complex data with the aid of decision vectors would seem appropriate. In contrast to data bank search procedures, classifiers require only limited storage space and the computation of the scalar products can be completed rapidly and without great expenditure. The spectra, absorbed by an on-line data system, might as a matter of routine be evaluated in a few seconds by means of several classifiers. In many instances, this preliminary interpretation would facilitate the evaluation of the mass spectra.

Acknowledgment

We express our graditude to Professor Dr. A. Maschka for his kind support in preparing this report. Mr. H. Urban deserves our thanks for his effective assistance with respect to the punching and programming work.

Bibliography

1. Benz, W., Massenspektrometrie organischer Verbindungen [Mass spectrometry of organic compounds], Academic Publishing Society, Frankfort/Main, 1969.

2. Buchanan, B.G., A.M. Duffield, A.V. Robertson, In: Mass spectrometry: Techniques and applications, p. 121, Publisher: W.A. Milne; New York: Wiley, 1971.

3. Crawford, L.R., J.D. Morrison, Anal. Chem., Vol. 40, p. 1469, 1968.

4. Crawford, L.R., J.D. Morrison, Anal. Chem., Vol. 43, p. 1790, 1971.

5. Grotch, S.L., Anal. Chem., Vol. 42, p. 1214, 1970.

6. Grotch, S.L., Anal. Chem., Vol. 43, p. 1362, 1971.

7. Isenhour, T.L., P.C. Jurs, Anal. Chem., Vol. 43, p. 20A, 1971.

8. Jurs, P.C., "Computerized learning machines for the interpretation of low-resolution mass spectrometry data," Dissertation, University of Washington, 1969, Dissertation Abstracts 30, No. 69-20241, 1969.

9. Jurs, P.C., Anal. Chem., Vol. 43, p. 22, 1971.

10. Jurs, P.C., B.R. Kowalski, T.L. Isenhour, Anal. Chem., Vol. 41, p. 21, 1969.

11. Jurs, P.C., B.R. Kowalski, T.L. Isenhour, C.N. Reilley, Anal. Chem., Vol. 41, p. 690, 1969.

12. Jurs, P.C., B.R. Kowalski, T.L. Isenhour, C.N. Reilley, Anal. Chem., Vol. 41, p. 1949, 1969.

13. Jurs, P.C., B.R. Kowalski, T.L. Isenhour, C.N. Reilley, Anal. Chem., Vol. 41, p. 1387, 1970.

14. Kowalski, B.R., C.F. Bender, Anal. Chem., Vol. 44, p. 1405, 1972.

15. Kowalski, B.R., C.F. Bender, J. Am. Chem. Soc., Vol. 94, p. 5632, 1972.

16. Kowalski, B.R., P.C. Jurs, T.L. Isenhour, C.N. Reilley, Anal. Chem., Vol. 41, p. 695, 1969.

17. Krenmayr, P., K. Varmuza, Allg. Prakt. Chem., Vienna, Vol. 23, p. 289, 1972.

18. Meyer-Brotz, G., J. Schurmann, Methoden der automatischen Zeichenerkennung [Methods of automatic pattern recognition], Munich: Oldenburg, 1970.

19. Nilsson, N.J., "Learning machines," New York: McGraw-Hill, 1965.

20. Stenhagen, E., S. Abrahamsson, F.W. McLafferty, Atlas of mass spectral data, New York: Intersci. Publ., 1969.

21. Wangen, L.E., T.L. Isenhour, Anal. Chem., Vol. 42, p. 737, 1970.